BIOINFORMATICS

Mikel Egaña Aranguren

www.sindominio.net/~pik

Manchester University Bioinformatics MSc

1.- INTRODUCCION:

- **Historia y bases generales de la Bioinformatica: bases de datos y algoritmos.**

2.- PROBLEMAS ACTUALES DE LA BIOINFORMATICA:

- **Informacion vs. conocimiento.**
- **Crecimiento de bases de datos.**
- **Brecha bioinformaticos/biologos.**

3.- ALGUNAS DE LAS ULTIMAS TENDENCIAS EN BIOINFORMATICA :

- **Hacia una computacion semantica: ontologias y la web semantica.**
- **Hacia una computacion distribuida: myGRID.**
- **Hacia una computacion estandar: los estandares en la investigacion con microarrays.**

1.- INTRODUCCION:

Attwood, T.K., Miller, C.J. "Bioinformatics goes back to the future". (2003). Nature Reviews. 4; 157-161.

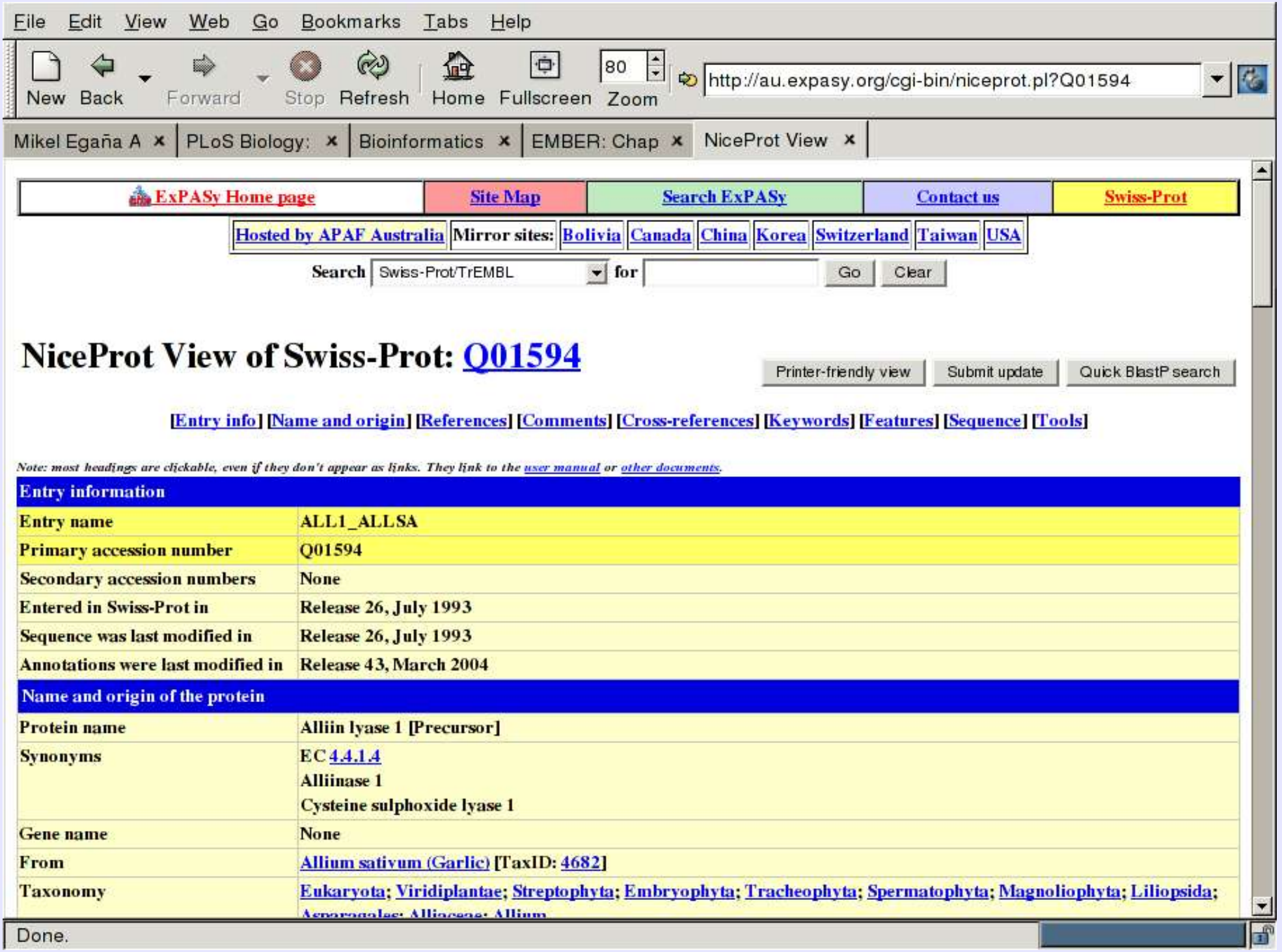
http://www.sindominio.net/~pik/bioinformatics.html

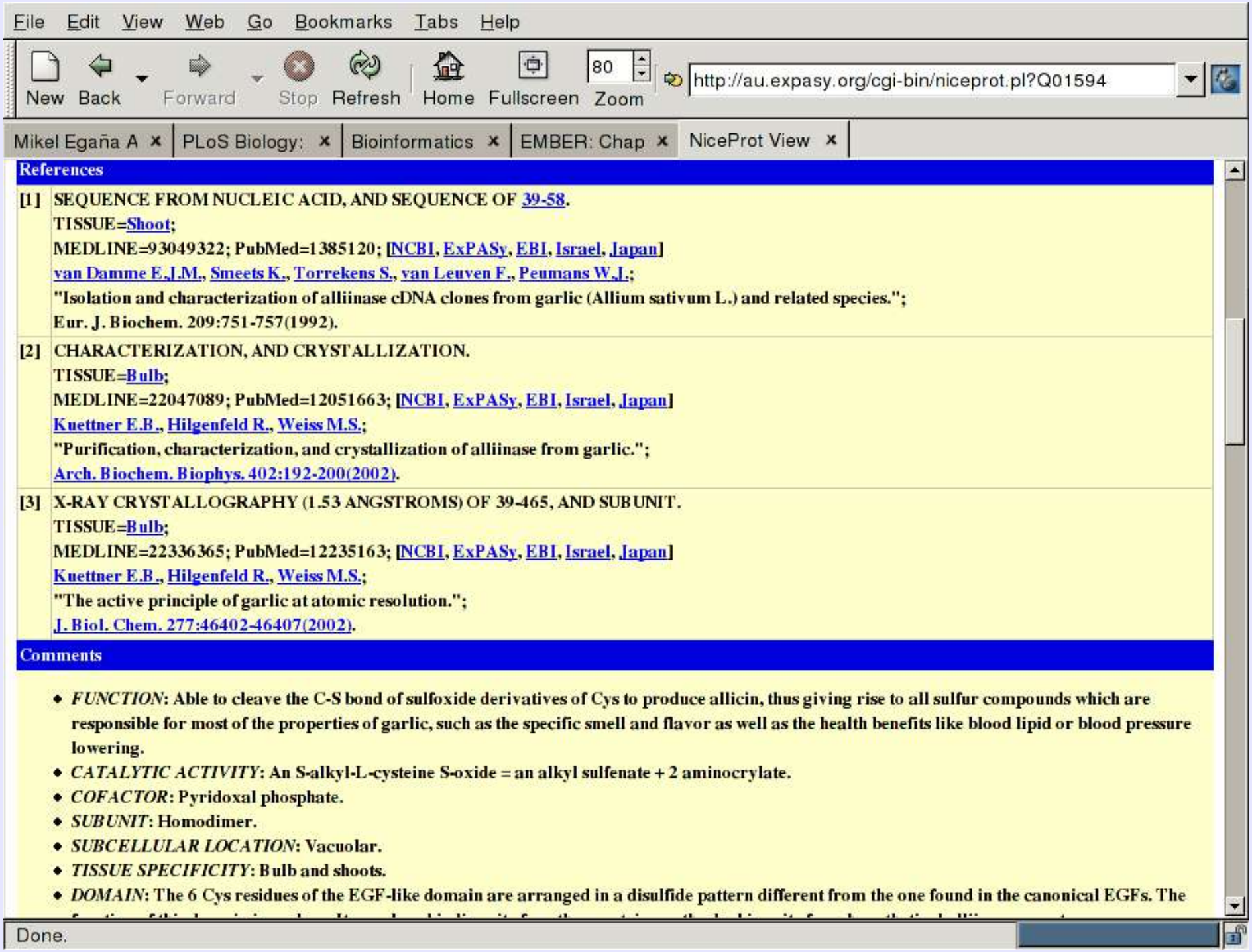
Biologia molecular en los 80: secuenciacion, estructuras.

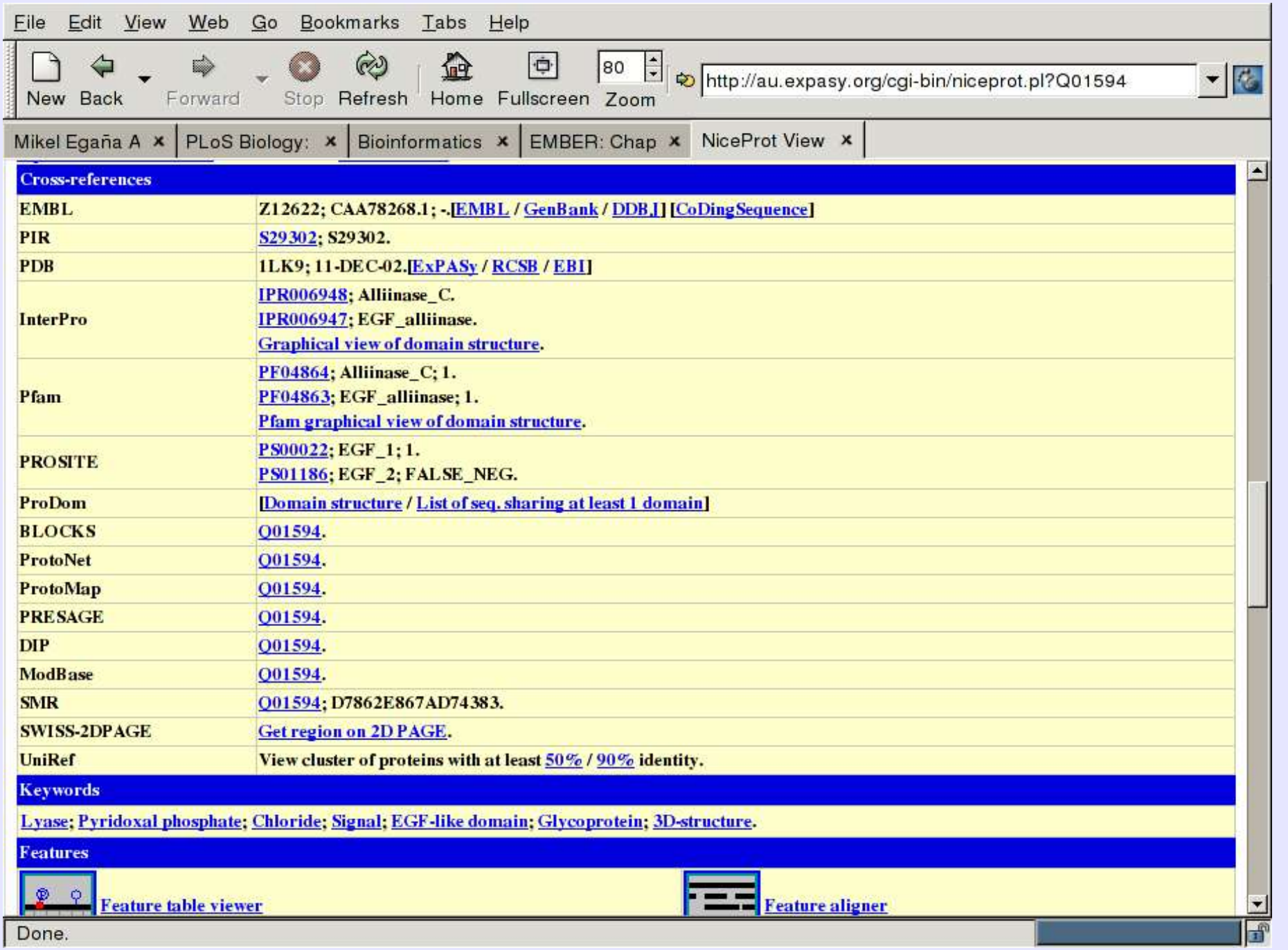
Centros principales:

- **Welcome Trust Genome Campus: EBI (European Bioinformatics Institute), Sanger Institute, Human Genome Mapping Project Resource Center.**
- **National Center for Biological Information (NCBI).**

Anotacion.


File Edit View Web Go Bookmarks Tabs Help
New Back Forward Stop Refresh Home Fullscreen Zoom
80
http://au.expasy.org/cgi-bin/niceprot.pl?Q01594
Mikel Egaña A
PLoS Biology:
Bioinformatics
EMBER: Chap
NiceProt View
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Hosted by APAF Australia
Mirror sites:
Bolivia
Canada
China
Korea
Switzerland
Taiwan
USA
Search
Swiss-Prot/TrEMBL
for
Go
Clear
NiceProt View of Swiss-Prot: Q01594
Printer-friendly view
Submit update
Quick BlastP search
[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]
Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.
Entry information
Entry name
ALL1_ALLSA
Primary accession number
Q01594
Secondary accession numbers
None
Entered in Swiss-Prot in
Release 26, July 1993
Sequence was last modified in
Release 26, July 1993
Annotations were last modified in
Release 43, March 2004
Name and origin of the protein
Protein name
Alliin lyase 1 [Precursor]
Synonyms
EC 4.4.1.4
Alliinase 1
Cysteine sulphoxide lyase 1
Gene name
None
From
Allium sativum (Garlic) [TaxID: 4682]
Taxonomy
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida;
Done.


File Edit View Web Go Bookmarks Tabs Help
New Back Forward Stop Refresh Home Fullscreen Zoom
80
http://au.expasy.org/cgi-bin/niceprot.pl?Q01594
Mikel Egaña A
PLoS Biology:
Bioinformatics
EMBER: Chap
NiceProt View
References
[1] SEQUENCE FROM NUCLEIC ACID, AND SEQUENCE OF 39-58.
TISSUE=Shoot;
MEDLINE=93049322; PubMed=1385120; [NCBI, ExPASy, EBI, Israel, Japan]
van Damme E.J.M., Smeets K., Torrekens S., van Leuven F., Peumans W.J.;
"Isolation and characterization of alliinase cDNA clones from garlic (Allium sativum L.) and related species.";
Eur. J. Biochem. 209:751-757(1992).
[2] CHARACTERIZATION, AND CRYSTALLIZATION.
TISSUE=Bulb;
MEDLINE=22047089; PubMed=12051663; [NCBI, ExPASy, EBI, Israel, Japan]
Kuettner E.B., Hilgenfeld R., Weiss M.S.;
"Purification, characterization, and crystallization of alliinase from garlic.";
Arch. Biochem. Biophys. 402:192-200(2002).
[3] X-RAY CRYSTALLOGRAPHY (1.53 ANGSTROMS) OF 39-465, AND SUBUNIT.
TISSUE=Bulb;
MEDLINE=22336365; PubMed=12235163; [NCBI, ExPASy, EBI, Israel, Japan]
Kuettner E.B., Hilgenfeld R., Weiss M.S.;
"The active principle of garlic at atomic resolution.";
J. Biol. Chem. 277:46402-46407(2002).
Comments
FUNCTION: Able to cleave the C-S bond of sulfoxide derivatives of Cys to produce allicin, thus giving rise to all sulfur compounds which are responsible for most of the properties of garlic, such as the specific smell and flavor as well as the health benefits like blood lipid or blood pressure lowering.
CATALYTIC ACTIVITY: An S-alkyl-L-cysteine S-oxide = an alkyl sulfenate + 2 aminocrylate.
COFACTOR: Pyridoxal phosphate.
SUBUNIT: Homodimer.
SUBCELLULAR LOCATION: Vacuolar.
TISSUE SPECIFICITY: Bulb and shoots.
DOMAIN: The 6 Cys residues of the EGF-like domain are arranged in a disulfide pattern different from the one found in the canonical EGFs. The
Done.


File Edit View Web Go Bookmarks Tabs Help
New Back Forward Stop Refresh Home Fullscreen Zoom
80
http://au.expasy.org/cgi-bin/niceprot.pl?Q01594
Mikel Egaña A
PLoS Biology:
Bioinformatics
EMBER: Chap
NiceProt View
Cross-references
EMBL
Z12622; CAA78268.1; -.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR
S29302; S29302.
PDB
1LK9; 11-DEC-02.[ExPASy / RCSB / EBI]
InterPro
IPR006948; Alliinase_C.
IPR006947; EGF_alliinase.
Graphical view of domain structure.
Pfam
PF04864; Alliinase_C; 1.
PF04863; EGF_alliinase; 1.
Pfam graphical view of domain structure.
PROSITE
PS00022; EGF_1; 1.
PS01186; EGF_2; FALSE_NEG.
ProDom
[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS
Q01594.
ProtoNet
Q01594.
ProtoMap
Q01594.
PRESAGE
Q01594.
DIP
Q01594.
ModBase
Q01594.
SMR
Q01594; D7862E867AD74383.
SWISS-2DPAGE
Get region on 2D PAGE.
UniRef
View cluster of proteins with at least 50% / 90% identity.
Keywords
Lyase; Pyridoxal phosphate; Chloride; Signal; EGF-like domain; Glycoprotein; 3D-structure.
Features
Feature table viewer
Feature aligner
Done.

File Edit View Web Go Bookmarks Tabs Help

New Back Forward Stop Refresh Home Fullscreen Zoom 80

http://au.expasy.org/cgi-bin/niceprot.pl?Q01594

Mikel Egaña A | PLoS Biology: | Bioinformatics | EMBER: Chap | NiceProt View

Features

Feature table viewer

Feature aligner

Key	From	To	Length	Description
SIGNAL	1	28	28	*Potential.*
PROPEP	29	38	10	
CHAIN	39	486	448	Alliin lyase 1.
DOMAIN	51	97	47	EGF-like (atypical).
BINDING	132	132		Chloride.
BINDING	136	136		Chloride.
BINDING	138	138		Chloride.
BINDING	289	289		Pyridoxal phosphate.
DISULFID	58	77		
DISULFID	79	88		
DISULFID	82	95		
DISULFID	406	414		
CARBOHYD	57	57		N-linked (GlcNAc...) (*Potential*).
CARBOHYD	184	184		N-linked (GlcNAc...).
CARBOHYD	229	229		N-linked (GlcNAc...) (*Potential*).
CARBOHYD	366	366		N-linked (GlcNAc...).

Sequence information

Length: 486 AA [This is the length of the unprocessed precursor]

Molecular weight: 55638 Da [This is the MW of the unprocessed precursor]

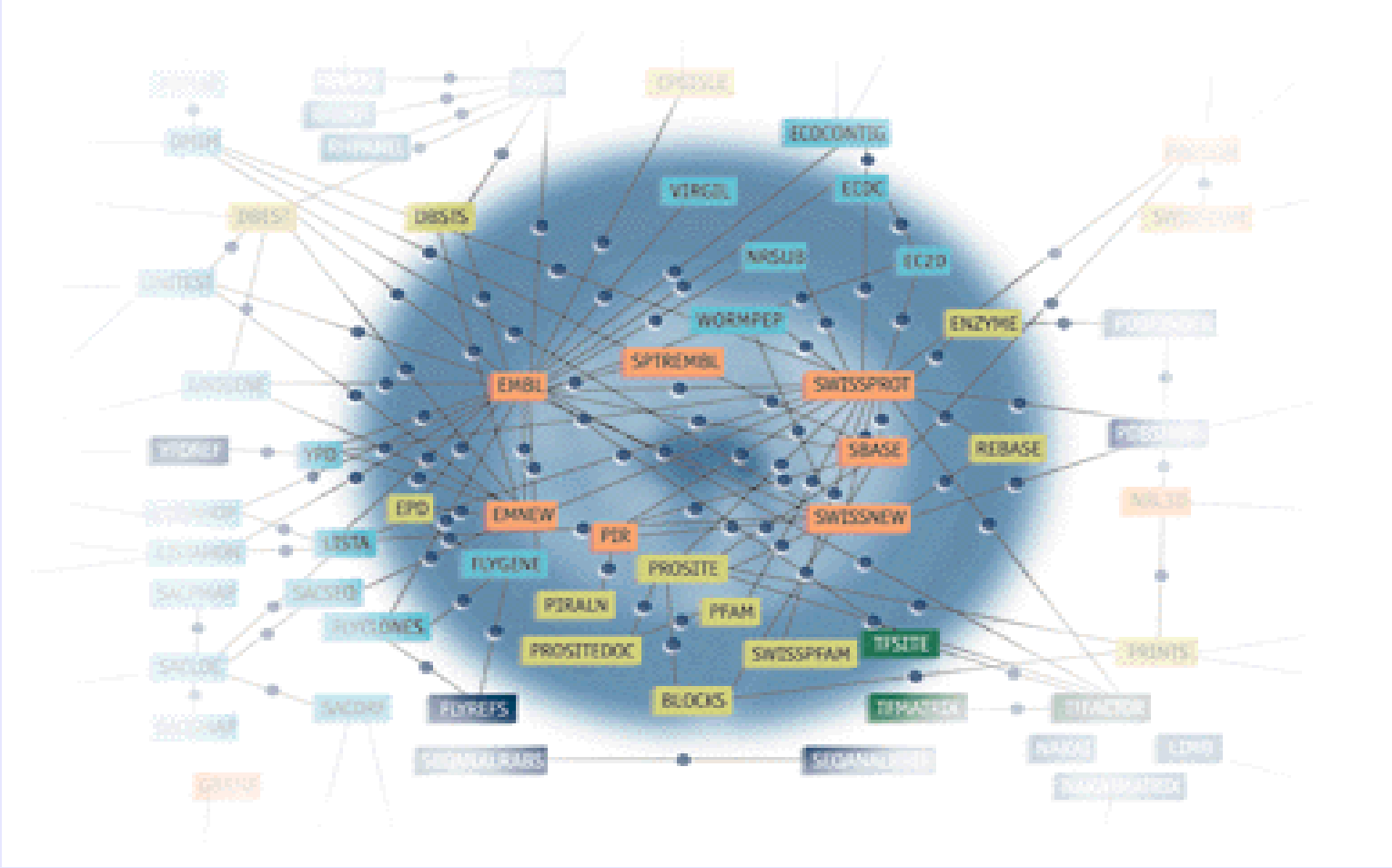
CRC64: D7862E867AD74383 [This is a checksum on the sequence]

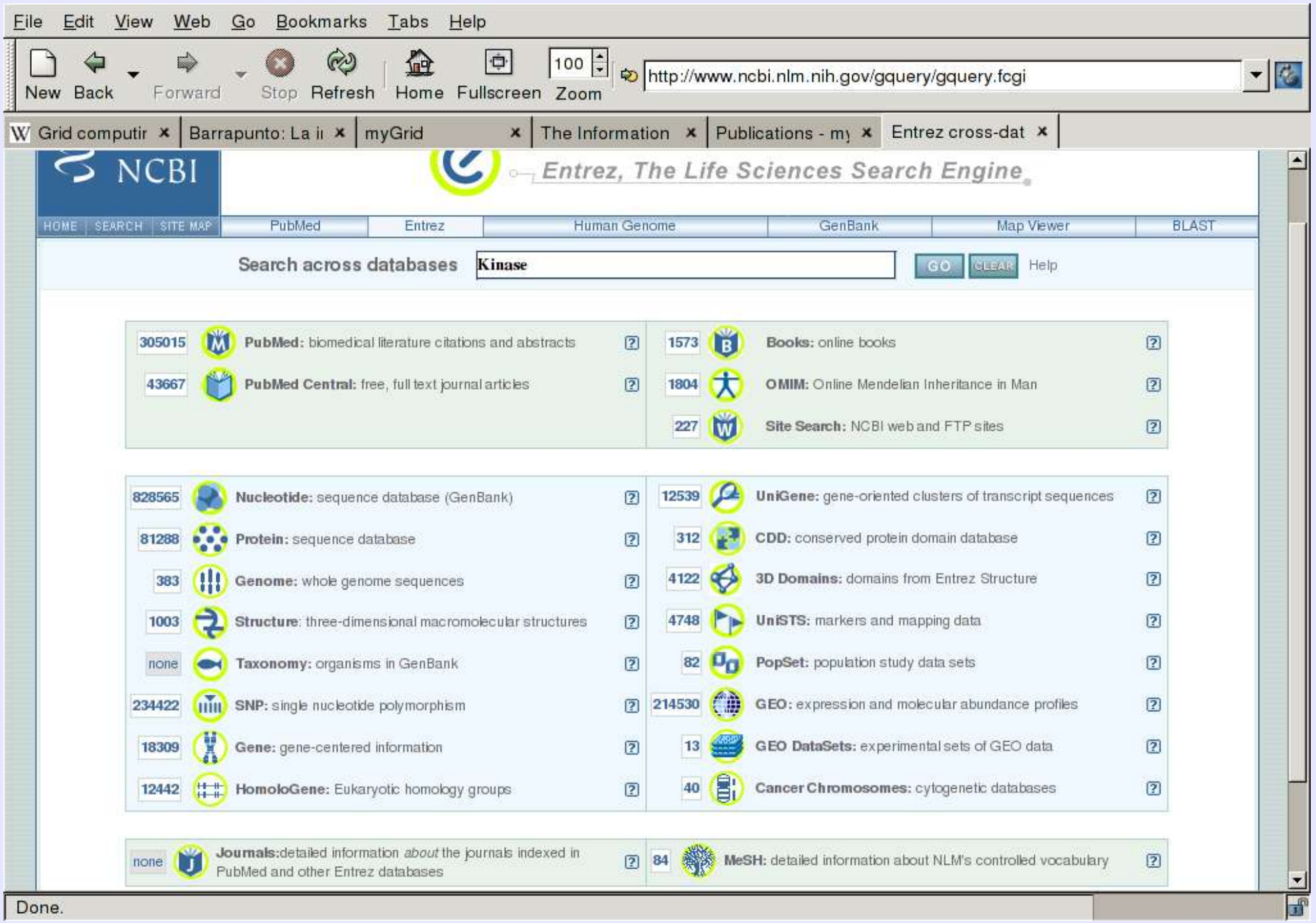
```
        10          20          30          40          50          60
         |           |           |           |           |           |
MVESYKKIGS CNKMPCLVIL TCIIMSNSLV NNNNMVQAKM TWTMKAAEEA EAVANINCSE

        70          80          90         100         110         120
```

Done.

Bases de datos:

- **Repositorios de secuencias: swiss-prot, trEMBL, GenBank.**
- **Sistemas que acceden a varias bd-s a la vez: SRS (EBI), Entrez (NCBI).**




File Edit View Web Go Bookmarks Tabs Help
New Back Forward Stop Refresh Home Fullscreen 100 Zoom
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
Grid computir
Barrapunto: La ii
myGrid
The Information
Publications - my
Entrez cross-dat
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed
Entrez
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
Kinase
GO
CLEAR
Help
305015 PubMed: biomedical literature citations and abstracts
43667 PubMed Central: free, full text journal articles
1573 Books: online books
1804 OMIM: Online Mendelian Inheritance in Man
227 Site Search: NCBI web and FTP sites
828565 Nucleotide: sequence database (GenBank)
81288 Protein: sequence database
383 Genome: whole genome sequences
1003 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
234422 SNP: single nucleotide polymorphism
18309 Gene: gene-centered information
12442 HomoloGene: Eukaryotic homology groups
12539 UniGene: gene-oriented clusters of transcript sequences
312 CDD: conserved protein domain database
4122 3D Domains: domains from Entrez Structure
4748 UniSTS: markers and mapping data
82 PopSet: population study data sets
214530 GEO: expression and molecular abundance profiles
13 GEO DataSets: experimental sets of GEO data
40 Cancer Chromosomes: cytogenetic databases
none Journals: detailed information about the journals indexed in PubMed and other Entrez databases
84 MeSH: detailed information about NLM's controlled vocabulary
Done.

Bases de datos de familias, patrones o motifs: abstraccion.

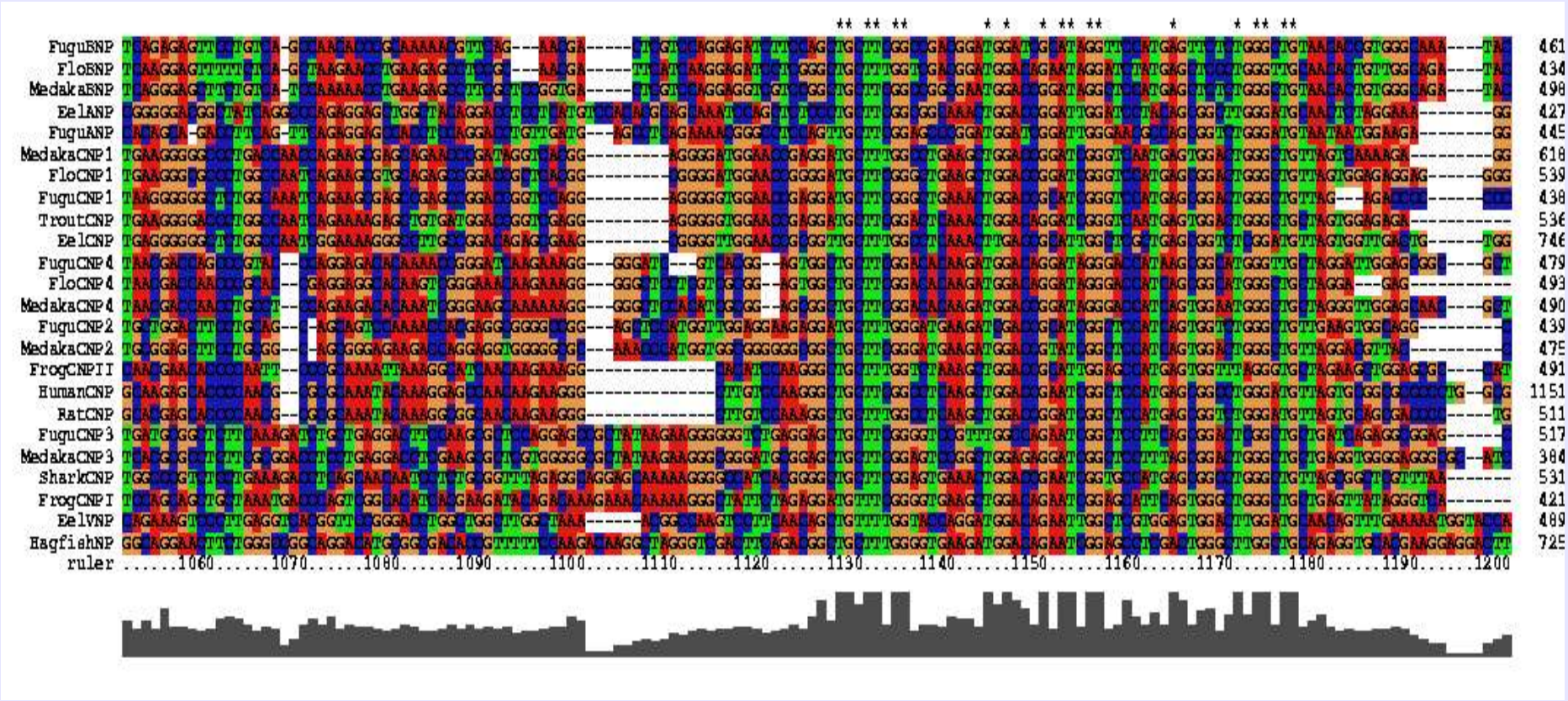
- **Mejor rendimiento en bajas similitudes.**
- **Busquedas funcionales.**
- **Unicos motifs: prosite, emotif. Expresiones regulares:**
  D-M-x-[ILV]-x{2}-G
- **Dominio: profiles (profiles), pfam (hidden markov models).**
- **Multiples motifs: fingerprints, blocks.**
- **Interpro.**

Bases de datos de estructuras, etc.

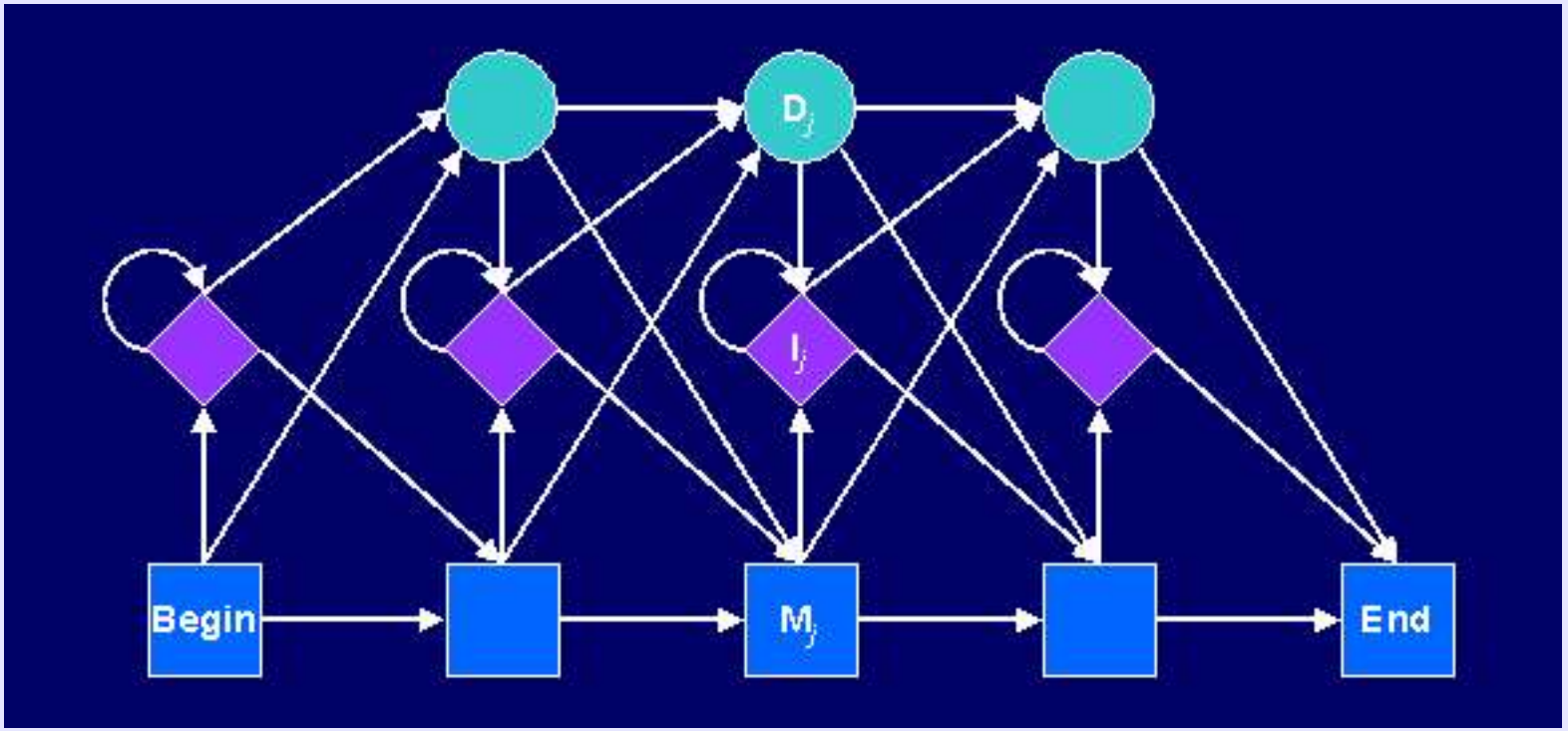
Algoritmos:

- **Prediccion de estructuras: CASP5 (Critical Assesment of technics for protein Structure Prediction): http://www.predictioncenter.llnl.gov/**

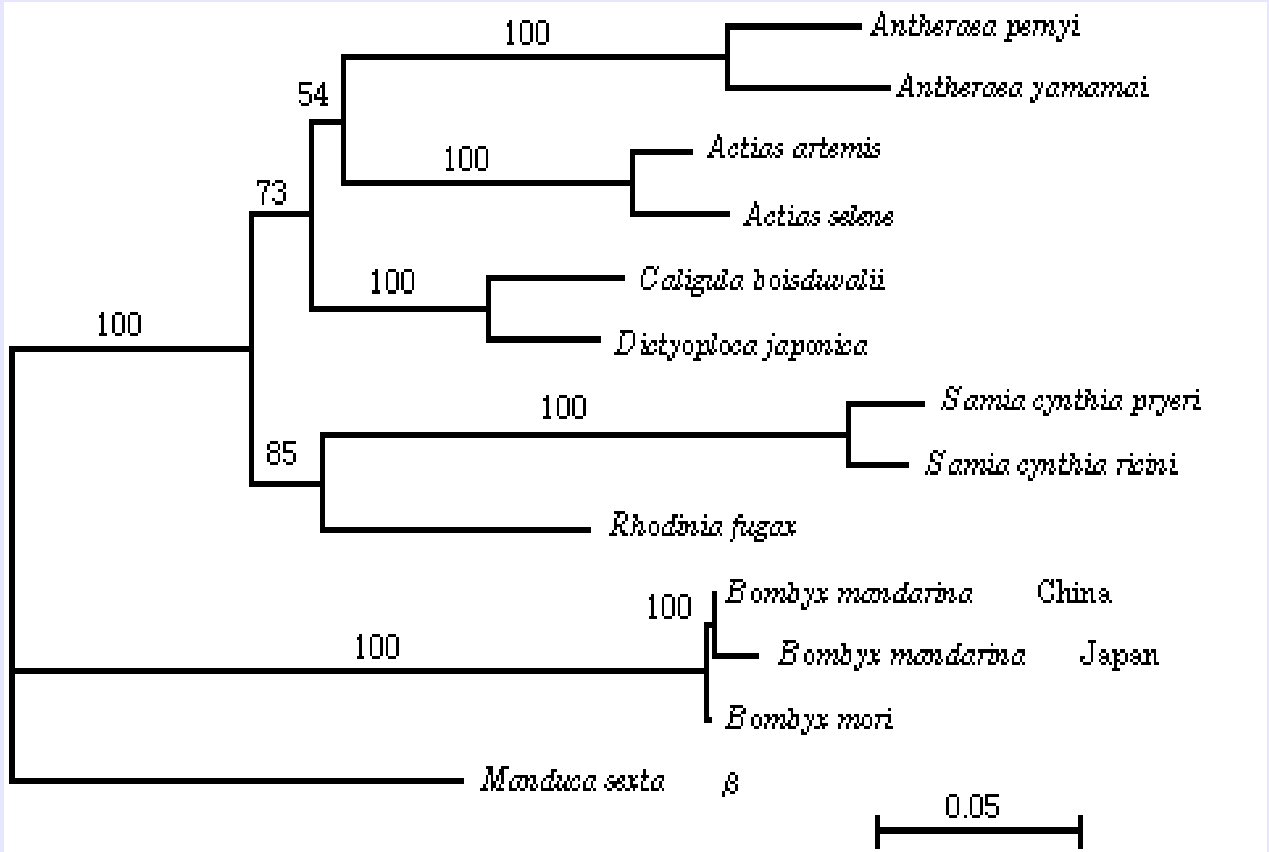
- **Alineamientos: algoritmos recursivos. Alineamiento local (BLAST: Basic Local Alignment Search Tool) o global. Matrices de similitud.**



- **Hidden Markov Models:**



- **Algoritmos filogeneticos:**



- **ETC ...**

- **BioPerl: http://www.bioperl.org.**
  "How Perl saved the human genome project":
  http://www.stanford.edu/class/gene211/handouts/How_Perl_HGP.html

2.- PROBLEMAS:

Informacion vs. conocimiento:

- **Infoberg: las bd-s duplican su contenido cada 9 meses.**
- **Anotacion: errores multiplicados, revisiones imposibles. Anotacion automatica. Similitud - funcion ???**
- **Homologos: ortologos - paralogos.**
- **Definicion de gen.**
- **Definicion de funcion.**
- **Naturaleza modular de las proteinas.**
- **Redundancia: NRDB: non-identical but redundant.**

Integracion de recursos: contingencia historica. N scripts son demasiados scripts para N herramientas.

Brecha bioinformaticos – biologos:

- **Falta de cultura computacional:**
  http://sindominio.net/biblioweb/telematica/command_es

3.-LINEAS DE FUTURO:

Hacia una computacion semantica:

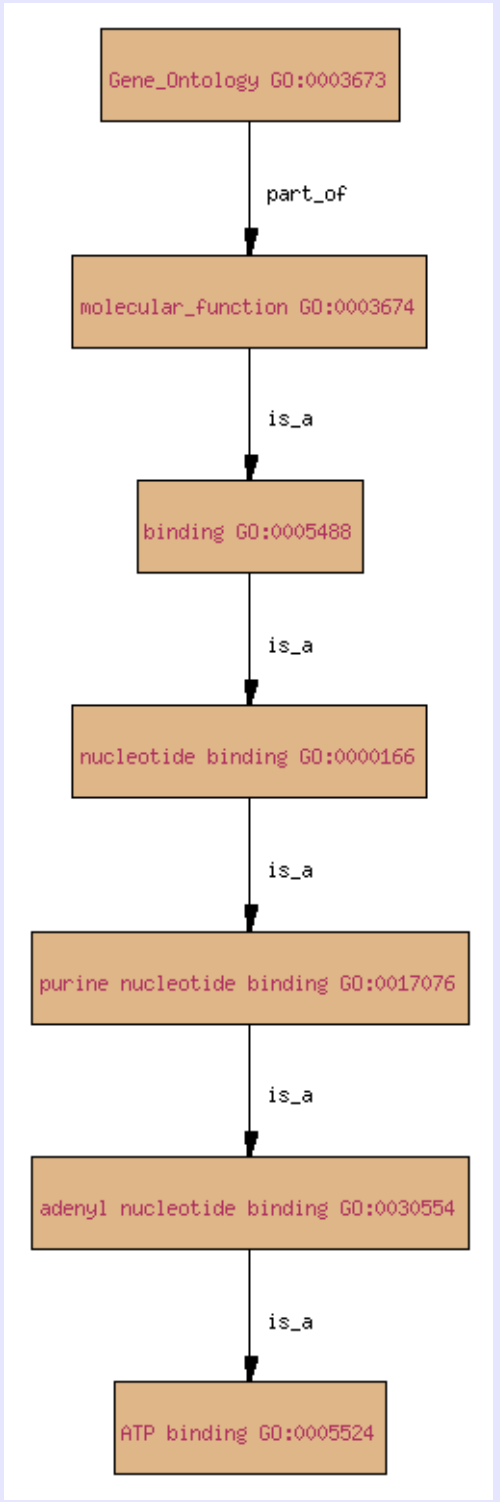
Ontologias: vocabulario + relaciones estructuradas:

Sistematizacion del conocimiento.

Semantica computacionalmente accesible.

Gene Ontology: proceso biologico, componente celular, funcion molecular.
http://www.geneontology.org

Open Biological Ontologies:
http://obo.sourceforge.net/



La web semantica:

- **W3C: http://www.w3.org/. World Wide Web Consortium.**
- **"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001.**
- **RDF (Resource Description Framework) + XML (eXtensible Markup Language).**
- **OWL (Web Ontology Language).**

Hacia una computacion distribuida:

- **Grid: computacion distribuida entre diferentes centros (con diferentes lenguajes de programacion, plataformas, ...) con una serie de protocolos estandar.**

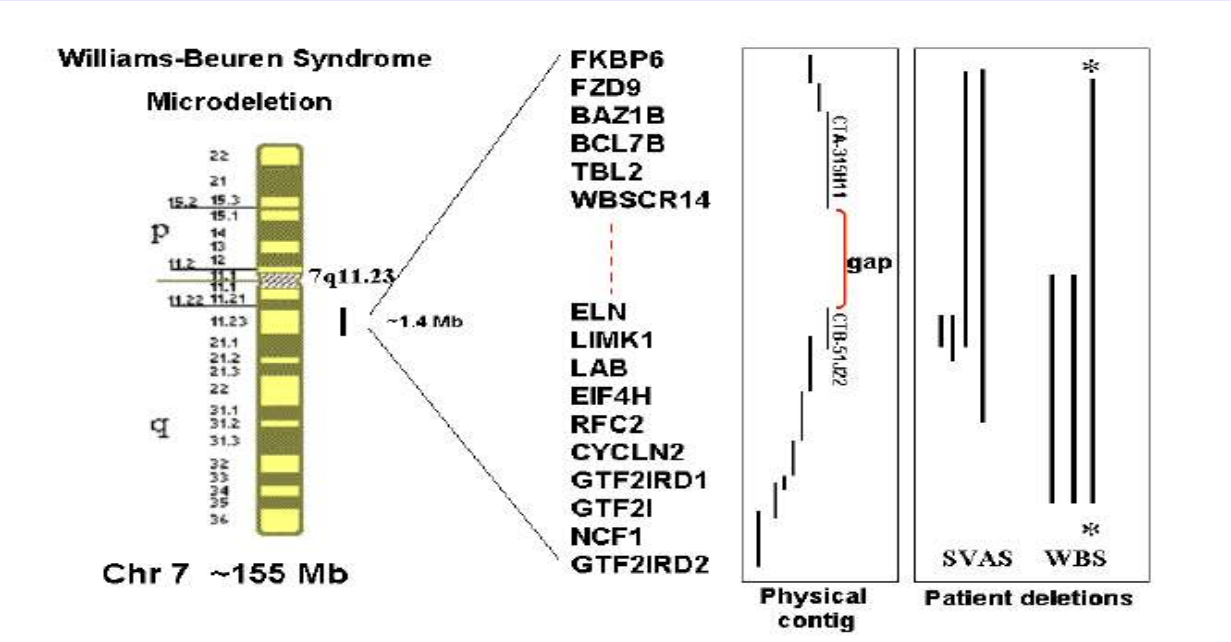
- **MyGrid: entorno virtual, automatizado y distribuido de experimentos in silico. http://www.mygrid.org.uk/**

- **Entorno comun de investigacion, automatico (rapidez), almacenaje de datos, distribuido y semantico, ...**

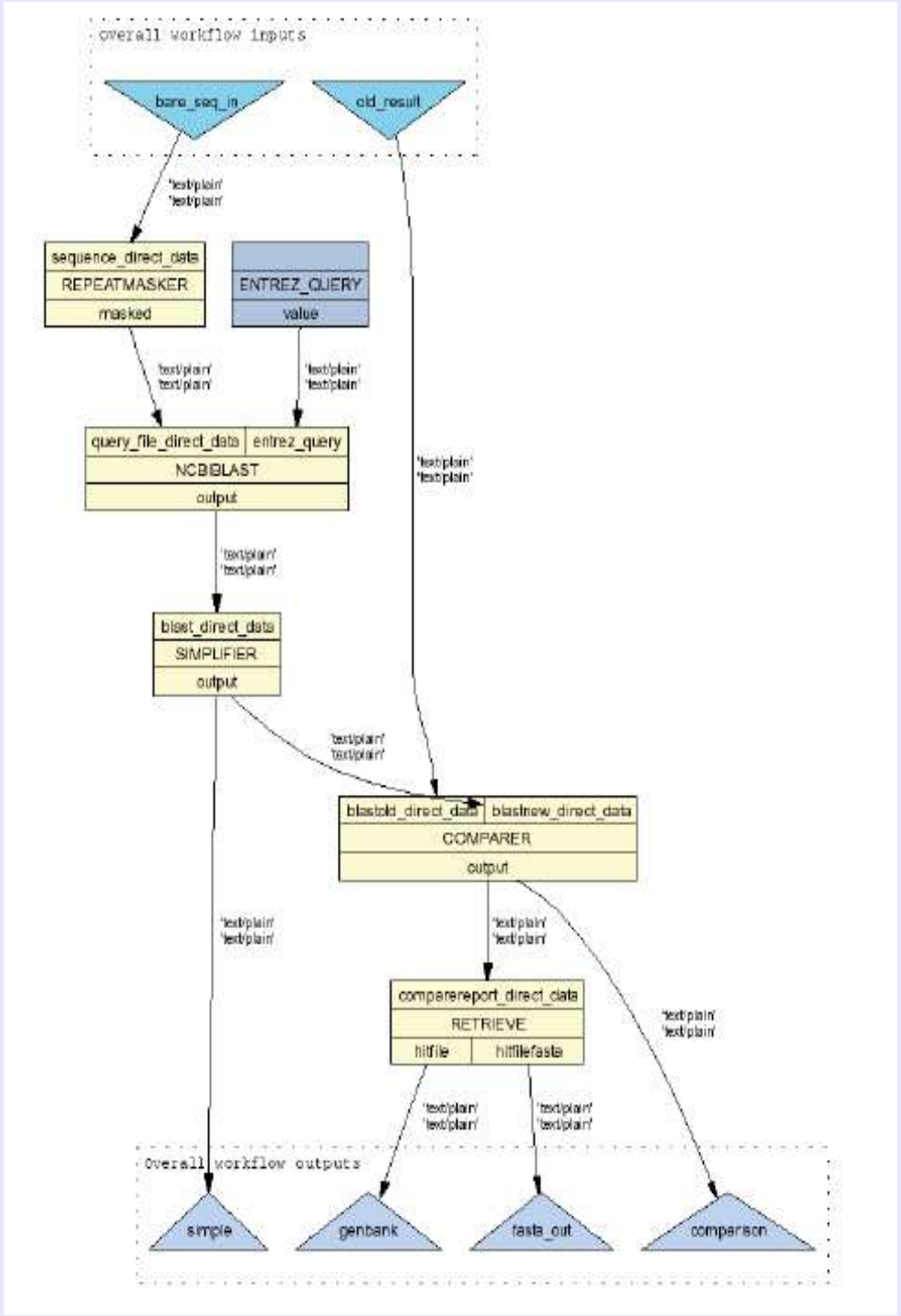
- **R. Stevens, H.J. Tipney, C. Wroe, T. Oinn, M. Senger, P. Lord, C.A. Goble, A. Brass and M. Tassabehji "Exploring Williams-Beuren Syndrome Using myGrid" to appear in Proceedings of 12th International Conference on Intelligent Systems in Molecular Biology, 31st Jul-4th Aug 2004, Glasgow, UK.**

Sindrome Williams – Beuren:

- **Deleciones cromosoma 7.**
- **Zona altamente repetitiva y compleja: WBS critical region, todavia sin caracterizar.**
- **Analisis periodicos >>> secuencias nuevas de esa zona >>> bateria de herramientas bioinformaticas >>> almacenaje de datos y procedimientos en el "LabBook".**

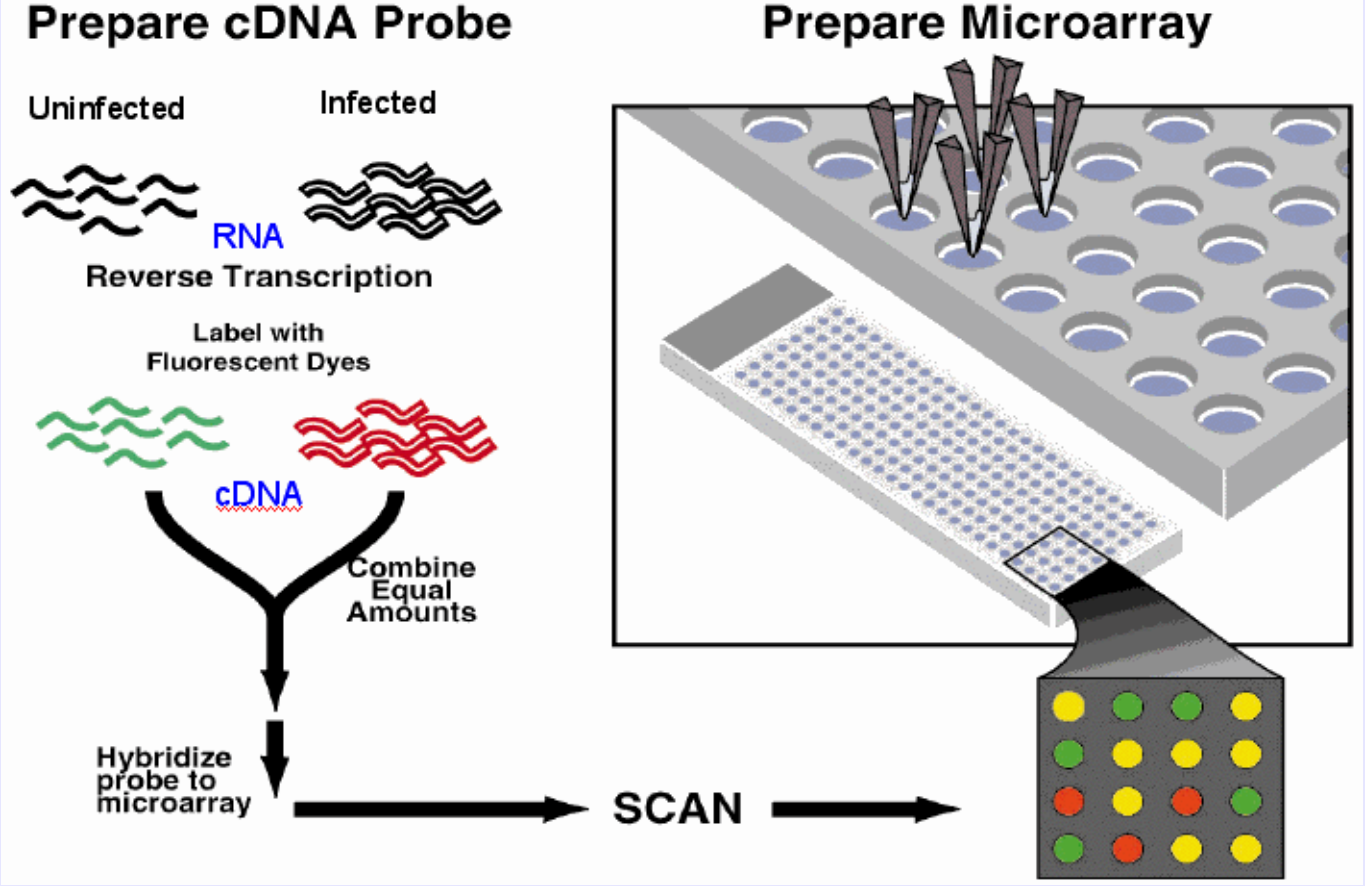


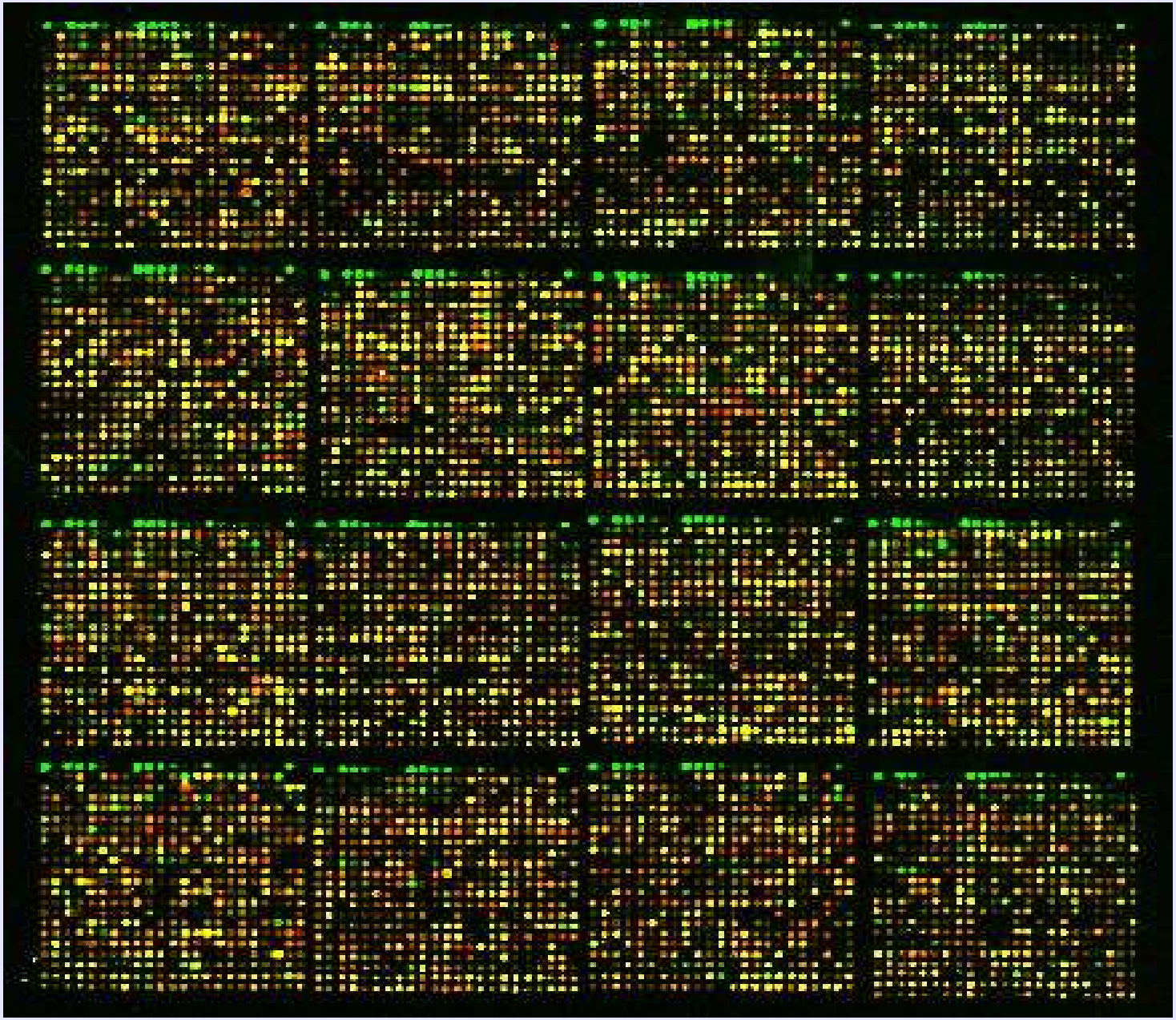
Workflows


overall workflow inputs
bare_seq_in
old_result
'text/plain'
'text/plain'
sequence_direct_data
REPEATMASKER
masked
ENTREZ_QUERY
value
'text/plain'
'text/plain'
'text/plain'
'text/plain'
query_file_direct_data
entrez_query
NCBIBLAST
output
'text/plain'
'text/plain'
'text/plain'
'text/plain'
blast_direct_data
SIMPLIFIER
output
'text/plain'
'text/plain'
blastold_direct_data
blastnew_direct_data
COMPARER
output
'text/plain'
'text/plain'
'text/plain'
'text/plain'
'text/plain'
'text/plain'
comparereport_direct_data
RETRIEVE
hitfile
hitfilefasta
'text/plain'
'text/plain'
'text/plain'
'text/plain'
Overall workflow outputs
simple
genbank
fasta_out
comparison

Hacia una computacion estandar:

- **Microarrays:**





Babelismo en bases de datos:

- **MGED: Microarray Gene Expresion Data: http://www.mged.org/. Estandares en el analisis, anotacion, almacenamiento y dispersion de datos sobre expresion de genes.**

- **MIAME: Minimal Information About a Microarray Experiment. Directivas para la anotacion de los resultados con microarrays. Adoptado por Nature, Cell. Doloroso para los biologos, bueno para la comunidad.**

- **MAGE-ML: Microarray Gene Expresion Markup Language.**

- **MAGE-OM: Microarray Gene Expresion Object Model.**